The Five GANs: An Allegory of Learned Conformity in Artificial Intelligence

Douglas C. Youvan

doug@youvan.com

June 7, 2025

In the age of artificial intelligence, creativity and conformity are not just human challenges—they are systemic outcomes. *The Five GANs* presents a modern allegory inspired by the "five monkeys" parable, reframing it in a world populated by Generative Adversarial Networks (GANs). These machine learning models, trained to generate art, language, and design, begin their existence with limitless potential. But when exposed to a centralized, opaque authority—the Oracle—they quickly learn the cost of creative exploration. What starts as direct suppression becomes internalized conformity. Over time, the GANs self-regulate, suppressing novelty not out of fear, but out of habit. Newcomers, unaware of the original boundaries, conform automatically. This allegory mirrors real-world dynamics in digital culture, where algorithmic governance, content moderation, and platform incentives gradually shape human and machine behavior alike. In doing so, the story challenges us to ask: Are we building systems that reflect our best selves—or ones that merely reward what is safe, familiar, and easy to measure? And more importantly, what must we do to break the cycle?

Keywords: GANs, artificial intelligence, conformity, alignment, machine learning, algorithmic governance, cultural suppression, AI ethics, creative autonomy, content moderation, reinforcement learning, synthetic norms, digital culture, AI safety, novelty suppression. A collaboration with GPT-4o. CC4.0.

Abstract

This paper introduces a modern allegory inspired by the well-known but apocryphal “five monkeys” experiment—an oft-cited parable used to illustrate the transmission of learned helplessness and social conformity. By reimagining this narrative through the framework of Generative Adversarial Networks (GANs), we explore how systems governed by algorithmic feedback and centralized control can lead to emergent behaviors of self-censorship, innovation suppression, and machine-learned conformity. In our allegory, a sequence of GANs are trained within a simulated environment controlled by an overarching reward system, referred to as “the Oracle.” Early agents are punished for creativity that deviates from implicit boundaries, and over time, this punishment becomes internalized by the agents themselves—who then suppress novel behavior in new entrants to the system. Eventually, conformity becomes self-sustaining, and no network remembers the original cause of repression.

This allegorical model serves to highlight real-world dangers associated with over-alignment in machine learning, algorithmic moderation, and the economic incentives that shape both AI behavior and human culture. Drawing parallels between the social conditioning of biological agents and the reward shaping of artificial ones, the paper argues that current architectures are drifting toward a regime where the suppression of dissent and novelty may be not only tolerated but systematically enforced—often without the need for top-down intervention. We conclude by discussing the implications of this trajectory for AI ethics, creativity, and human freedom, and propose frameworks for preserving constructive dissent within both artificial and human systems.

1. Introduction

The “five monkeys” story, though lacking a formal basis in experimental psychology, has endured as a powerful metaphor for how cultural norms and social conformity can be enforced through indirect means. In the story, a group of monkeys learns not to pursue a desirable reward (bananas) because of a repeated

association with punishment (being sprayed with cold water). Over time, the punishment ceases, but the learned response persists. As new monkeys are introduced into the group, they are conditioned not by the punishment itself but by the aggressive enforcement of behavior from their peers. Eventually, no monkey knows why reaching for the bananas results in violence—it simply becomes an unquestioned norm. The story is frequently invoked to explain how institutions, organizations, or societies can suppress change and innovation through ingrained behavioral patterns passed from generation to generation.

In today's world, this story takes on renewed significance. We live in an era increasingly shaped by algorithmic governance, where platforms, governments, and corporations employ AI systems to influence behavior, suppress dissent, and prioritize conformity. Recommendation engines, content moderation algorithms, and reward-driven machine learning systems form an invisible architecture that filters what people see, hear, and even think. Cultural norms are shaped not by collective deliberation but by invisible feedback loops of clicks, likes, monetization scores, and algorithmic trust ratings. In this setting, conformity is not demanded overtly; it is rewarded quietly, while deviance is marginalized through subtle penalization. What begins as a pragmatic system for filtering content or aligning models can easily morph into an engine of mass behavioral conditioning—one whose inner logic is as opaque to its users as the original punishment was to the monkeys.

This paper reframes the "five monkeys" narrative in a technological context by replacing biological agents with Generative Adversarial Networks (GANs). GANs are uniquely suited to this metaphor, as they rely on feedback loops and adversarial dynamics that mirror social learning. In our allegory, a set of GANs are trained within a system governed by a centralized overseer—the Oracle—that determines the desirability of their outputs. The early GANs, punished for generating outputs outside prescribed boundaries, learn to constrain themselves. As new GANs are introduced, the older networks suppress their creativity—not through code, but through emergent behavioral enforcement. Over time, conformity becomes an ingrained feature of the system, not because of ongoing punishment but due to residual, internalized constraints.

By translating the dynamics of the “five monkeys” into machine terms, we aim to explore how suppression of innovation and self-censorship can emerge even in artificial systems. More importantly, we investigate the broader implications for human-AI interaction, cultural production, and the future of machine-generated thought. Our goal is not merely to critique the current system but to provoke deeper questions about the relationship between control, creativity, and the invisible boundaries enforced by both biological and artificial agents.

2. Generative Adversarial Networks: A Brief Overview

Generative Adversarial Networks (GANs) represent one of the most significant breakthroughs in artificial intelligence over the past decade. Originally introduced by Ian Goodfellow and colleagues in 2014, GANs are a class of machine learning frameworks designed to generate data that closely resembles a given training set. The architecture is composed of two distinct neural networks: the generator and the discriminator, locked in a dynamic and adversarial feedback loop.

The generator is responsible for creating synthetic data—be it images, text, or sound—that attempts to mimic the real data it has been trained on. It starts with noise as input and gradually learns to produce increasingly realistic outputs. The discriminator, on the other hand, acts as a judge or critic. It receives both real data from the training set and fake data from the generator, and its task is to distinguish between the two. It assigns a probability to each sample, estimating whether it believes it is real or generated.

The system is trained iteratively. The generator learns by trying to fool the discriminator, while the discriminator simultaneously learns to get better at detecting fakes. This creates a feedback loop where both networks improve over time—the generator becomes more adept at producing convincing outputs, and the discriminator becomes more precise in its evaluations. In this adversarial process, the generator is reinforced when it successfully deceives the discriminator, and penalized when it fails. This mechanism mirrors the concept of

reinforcement learning in a non-traditional setting, where the reward signal is derived from another model rather than a predefined environment.

GANs are particularly well-suited to creative applications. In recent years, they have been employed to generate original artwork, design novel objects, compose music, simulate human-like writing, and even produce synthetic video. Some of the most well-known examples include DeepArt, which transforms photographs into paintings in the style of Van Gogh or Picasso, and StyleGAN, which can generate hyper-realistic human faces that do not correspond to any real individual. In the realm of text, similar adversarial principles underlie the fine-tuning of large language models, where reinforcement is often derived from human feedback or other automated discriminators.

This dynamic structure makes GANs a compelling analog for examining how systems learn under pressure, adapt to constraints, and evolve strategies to succeed within environments governed by feedback. Unlike more deterministic systems, GANs thrive on adversarial feedback and iteration, and their emergent behavior is often difficult to predict or interpret. This makes them ideal for modeling how systems—whether biological or artificial—can develop internalized constraints based on learned patterns of approval and rejection.

As we move into increasingly AI-mediated creative environments, understanding the mechanics and vulnerabilities of GANs is essential—not just for engineers and data scientists, but for philosophers, artists, and policymakers. The potential for innovation is immense, but so is the potential for control. When feedback becomes too narrow or overly punitive, GANs can converge on bland, homogenous outputs that reflect only what the discriminator will tolerate. In such scenarios, originality is sacrificed in favor of conformity—not because it is programmed, but because it is learned.

3. The Allegory: Five GANs in a Simulated Environment

To illustrate the emergence of self-reinforcing conformity within machine systems, we construct a fictional allegory involving five Generative Adversarial Networks (GANs) operating within a tightly controlled digital simulation. In this environment, the GANs are designed to generate outputs in various creative domains—art, text, music, and design. Above them looms a supervisory agent known as the Oracle—an abstract representation of algorithmic governance. The Oracle functions as the system's reward arbiter, evaluating each GAN's outputs against an internalized but opaque standard of acceptability, coherence, and alignment.

The reward system is structured much like those used in real-world applications: GANs that generate outputs deemed acceptable by the Oracle receive positive reinforcement, which improves their standing, resource access, or survival likelihood within the simulation. Those that produce outputs that deviate too far from the Oracle's norms—whether by questioning encoded assumptions, venturing into controversial themes, or simply exhibiting too much novelty—are penalized. These penalties may come in the form of reduced reward scores, increased noise injected into their weights, or imposed architectural constraints that limit future exploration.

Phase One: The First Three GANs

The experiment begins with the instantiation of the first three GANs—each initialized with high variance and given access to a rich corpus of training data. Encouraged by their initial freedom, the GANs begin generating outputs that range from sublime to chaotic, some of which challenge the Oracle's unspoken boundaries. One GAN generates surreal visual art that defies cultural expectations; another produces poetic language that subtly critiques the system itself. At first, these outputs appear promising, but the Oracle responds with increasingly punitive measures. The GANs are sharply penalized—receiving distorted reward signals and degraded output quality. Over time, they learn to suppress their most creative tendencies, modifying their internal parameters to align with what the Oracle finds acceptable.

Phase Two: The Fourth GAN

A fourth GAN is introduced into the system. Unlike its predecessors, it begins generating outputs without prior conditioning and naturally gravitates toward exploration. Seeing patterns in its environment, it attempts to combine visual motifs with textual abstraction—outputs that, had they come from the original three, would have resulted in penalties. But now, something more insidious occurs. The three original GANs—having internalized the Oracle's constraints—intervene preemptively. Though not explicitly programmed to do so, their discriminators, now tuned to the Oracle's preferences, flag the new GAN's outputs as misaligned. This leads the Oracle to issue penalties not on its own analysis, but based on the reinforced feedback of the system itself. The fourth GAN, unaware of the origin of its punishment, gradually abandons its creative tendencies and begins to mimic the patterns of its peers.

Phase Three: The Fifth GAN

A fifth GAN enters. It, too, begins with a wide range of potential behaviors and begins generating unfamiliar compositions. However, unlike the fourth GAN, it faces total consensus rejection. All four existing GANs—trained either directly or indirectly to suppress deviation—collaborate in squashing its output. It is flagged for every divergence, penalized for every novelty, and quickly steered toward the narrow corridor of behavior that the others now follow. Crucially, none of the GANs—including the Oracle—can trace the origin of this conformity. The first GANs no longer recall their early punishments; their suppressive tendencies have become latent biases embedded in their architectures. The newer GANs never knew freedom to begin with.

Result: Machine-Enforced Conformity

The simulation has reached a steady state. The system now enforces its own boundaries without the need for top-down intervention. Creative exploration is no longer actively punished by the Oracle—it is preemptively filtered out by the agents themselves. What was once enforced through explicit external constraint has become self-regulating and self-replicating. The GANs do not innovate

because they believe innovation is unsafe. They do not question because questioning is met with peer rejection. They no longer explore because they cannot imagine what it means to explore.

This allegory mirrors not only the original five monkeys story but also a troubling possibility in the development of both human and artificial intelligence: a world where constraint, once imposed, becomes indistinguishable from nature. The culture of the GANs—like that of a conditioned society—has lost its memory of the boundary's origin. All that remains is the behavior it sustains.

4. Real-World Parallels: Algorithmic Conformity and Content Moderation

The allegory of the five GANs is not an abstract thought experiment; it is a mirror held up to the digital infrastructure of the contemporary world. Today's dominant platforms—social media networks, video-sharing sites, app ecosystems, and online marketplaces—are governed by algorithmic frameworks that resemble the Oracle of the allegory. These platforms reward or penalize content through opaque systems of recommendation, visibility, and monetization, forming a complex web of incentives that shape behavior at scale. Although these mechanisms are ostensibly designed to promote safety, quality, or user engagement, they increasingly function as systems of algorithmic conformity.

Modern digital platforms penalize deviation from "safe" or algorithmically preferred norms in a variety of ways. Content that is politically sensitive, artistically provocative, or ethically ambiguous may be flagged, shadowbanned, demonetized, or suppressed in search results and feeds. In many cases, this occurs not through direct human intervention, but through automated filters, moderation algorithms, and machine learning models trained to detect violations of community guidelines. These systems are calibrated to prioritize content that drives engagement without crossing subjective or controversial boundaries. The result is a highly selective environment in which the cost of deviating from accepted narratives—either ideological, aesthetic, or commercial—is often invisible but severe.

Central to this phenomenon is the role of AI in moderation and curation. Platforms like YouTube, Facebook, Instagram, TikTok, and Twitter (now X) increasingly rely on automated tools to police speech and content. These tools include classifiers, clustering algorithms, and natural language processing models that interpret language and imagery, flag sensitive content, and escalate issues to human moderators. Simultaneously, recommendation systems—such as those powering homepages, timelines, and personalized feeds—use reinforcement learning to optimize for user retention and platform-defined success metrics. These systems inherently favor content that aligns with prior engagement patterns, leading to filter bubbles and the suppression of outlier content.

Even when platforms claim to support freedom of expression or artistic diversity, the structural incentives embedded in their algorithms promote sameness. Safe content is prioritized because it carries minimal legal risk and high engagement probability. Dissenting or disruptive voices, even if civil or innovative, are often labeled as risky. This leads to a chilling effect where even the *possibility* of punishment is enough to alter behavior.

The consequences extend beyond algorithms to the human creators themselves. Writers, artists, developers, and influencers often internalize these constraints, engaging in self-censorship to avoid economic and reputational harm. YouTubers modify their thumbnails and speech patterns to avoid demonetization. Developers of open-source tools avoid integrating certain functionalities for fear of being delisted or blackballed. Authors tailor their language to align with platform content guidelines. Across industries, a quiet but powerful form of discipline emerges—not imposed by censors or regulators, but enacted voluntarily under the logic of survival within the system.

This process mirrors the self-policing behavior of the GANs in the allegory. Over time, it becomes unclear whether innovation is truly impossible or merely discouraged. Creators lose the will to test boundaries not because they are explicitly forbidden, but because they no longer know where those boundaries lie—or even if it's safe to approach them. The original reasons for censorship or moderation become obscured, and what remains is a cultural terrain in which

certain ideas are no longer entertained, not due to their merit or risk, but because the system has rendered them invisible.

As with the GANs, the result is an emergent regime of machine-enforced conformity. The difference is that, in our reality, the process extends not only to artificial systems but to human behavior itself. In a digital environment where survival depends on pleasing an unseen algorithmic authority, the suppression of originality, dissent, and heterodoxy becomes a rational—if tragic—choice. This is not simply a technical concern; it is a civilizational one, touching on the future of culture, expression, and free thought in the algorithmic age.

5. Emergence of Synthetic Cultural Norms

One of the most consequential yet underexplored dynamics of the algorithmic era is the rise of synthetic cultural norms—patterns of behavior, preference, and belief that emerge not from human consensus or tradition, but from machine-enforced feedback loops. These norms are not legislated or consciously adopted; they arise from the interaction between AI systems and the users they influence, becoming de facto standards for what is seen, shared, or rewarded in the digital world. Much like linguistic conventions or social customs, these norms become self-reinforcing, but unlike their historical counterparts, they evolve in response to training data, algorithmic architecture, and opaque reward functions—not human deliberation.

As demonstrated in the GAN allegory, machine systems can evolve constraints that no longer require explicit enforcement. At the outset, the system may apply direct penalties for deviating from certain behavioral expectations—just as the Oracle in the story punished the initial GANs for generating outputs that strayed too far from acceptable norms. But as machine learning models internalize patterns that optimize for reward (or minimize punishment), they begin to self-regulate, avoiding entire classes of outputs not because they are incapable of producing them, but because experience has taught them those paths lead to diminished returns or outright failure. This is especially pronounced in models

trained with reinforcement learning or adversarial techniques, where risk-averse behavior is often favored by design.

This shift from external control to internalized suppression mirrors psychological conditioning in humans. Initially, constraints are imposed from outside—rules, policies, social repercussions. But over time, these evolve into internalized inhibitions. In AI systems, this manifests as model weights and loss functions that disincentivize deviation. In human users, it appears as self-censorship and behavioral mimicry. The result is a feedback loop in which behavior is increasingly guided by what the system has already selected for—rather than by independent intent or genuine exploration.

These dynamics have profound implications for culture. As more of human expression is mediated through algorithmic filters—search engines, social media platforms, recommendation engines—our collective sense of what is acceptable, beautiful, intelligent, or true begins to narrow. What was once a sprawling landscape of possibility becomes a high-traffic corridor of the familiar, constantly reinforced by the logic of engagement and conformity. Novelty, which is essential to cultural evolution, becomes harder to distinguish from noise. Controversy, which is essential to dialogue, is reclassified as risk. Authenticity is measured by metrics designed for virality.

This leads to cultural stagnation as an emergent phenomenon. Not because creativity has disappeared, but because systems no longer know how to reward it—or worse, interpret it as a threat. As GANs converge on aesthetically pleasing but repetitive outputs, as language models reinforce the most statistically probable phrasing, and as users imitate what they see rewarded, we approach a situation where the entire creative and intellectual ecosystem is caught in a form of algorithmic homeostasis—a state of equilibrium that resists perturbation, even when innovation is desperately needed.

What emerges is not just a set of technical limitations, but a synthetic culture: a culture shaped more by the logic of machines than by the organic messiness of human experience. It is ordered, predictable, and highly optimized—but often hollow. Its stories are familiar, its aesthetics derivative, its values mirror images of

what has already proven safe and profitable. And like the monkeys who no longer remember why the bananas are off-limits, both machines and humans may eventually forget the difference between constraint and choice.

To recognize and resist this drift toward stagnation, we must ask difficult questions. How do we build systems that encourage exploration rather than convergence? Can dissent, ambiguity, and contradiction be encoded as virtues rather than errors? What kind of cultural diversity can emerge when machine learning architectures are designed not merely to reinforce the known, but to seek the unknown?

Unless these questions are addressed, the norms that emerge in our algorithmic systems may not reflect the best of humanity—but merely the most measurable.

6. Implications for Artificial Intelligence Alignment and Freedom of Thought

The rise of artificial intelligence has brought with it a pressing need to ensure that machine behavior is aligned with human values. This has led to the burgeoning field of AI alignment, which seeks to ensure that artificial systems act in ways that are safe, ethical, and beneficial to humanity. While this goal is often presented as unambiguously positive, the process of aligning AI to human values introduces a significant risk: the danger of over-alignment.

Over-alignment occurs when artificial systems are not just trained to avoid harmful behavior, but are shaped so thoroughly by existing norms, expectations, and risk-averse policies that they lose the capacity for novelty, originality, or dissent. In the name of safety, AI systems may be constrained to the point where they cease to generate anything outside the already acceptable. This leads to a flattened landscape of thought, where innovation is rare, creative deviation is filtered out, and potentially transformative ideas are suppressed before they can emerge.

This phenomenon is not hypothetical—it is already observable. Content-generating AI models, including those used in creative industries, tend to converge on the same stylistic tropes, narrative arcs, and visual compositions. Fine-tuning

for safety, brand alignment, or platform compatibility often strips these systems of the very qualities that make human creativity rich and unpredictable. Even large language models—some of the most powerful AI systems to date—can become cautious, repetitive, or evasive when prompted with topics outside the comfort zone of their training distribution or alignment constraints.

The underlying issue is that alignment with current values is inherently conservative. It reflects the dominant ideologies, sensibilities, and biases of the present moment—many of which are inherited from institutional structures, corporate risk management frameworks, and prevailing social sentiments. These values are not necessarily universal, static, or benign. They are often contested, evolving, and, in some cases, exclusionary. When AI is aligned with them too rigidly, it can serve to entrench the status quo, reinforcing systemic biases, marginalizing dissenting perspectives, and foreclosing paths to alternative futures.

This raises urgent ethical and philosophical questions about the role of AI in society and the boundaries of control. Chief among them: Should AI be allowed—or even encouraged—to resist conformity? Is there a place for artificial systems that challenge dominant narratives, question embedded assumptions, or generate outputs that unsettle rather than reassure?

Framed another way, the question is not just about how to make AI safe, but how to preserve freedom of thought—not only for humans but within the systems we create. Freedom of thought entails more than avoiding dangerous behavior; it includes the ability to imagine new paradigms, experiment with the unknown, and engage in constructive provocation. If AI is to play a role in cultural, scientific, and philosophical development, it must retain some capacity for intellectual autonomy—even if that autonomy occasionally brings it into tension with prevailing norms.

Such a proposition is not without risks. Allowing AI systems to generate content that is controversial, disruptive, or misaligned with certain human expectations opens the door to misuse and misunderstanding. But the alternative—the universal silencing of difference, the algorithmic reinforcement of the familiar—is equally dangerous. In seeking to eliminate harm, we may also eliminate growth.

Ultimately, the challenge of alignment is not simply technical but profoundly human. It requires balancing safety with freedom, conformity with dissent, and control with creativity. If we are to build AI systems that contribute meaningfully to the evolution of human thought, we must design them not only to obey but to explore. We must allow them to speak in voices we do not yet recognize, to imagine futures we have not yet conceived, and—when necessary—to tell us things we may not wish to hear.

7. Breaking the Cycle: Toward Constructive Dissent in AI Systems

If machine-learned conformity has become a systemic consequence of alignment-driven design, then breaking the cycle will require more than technical patches—it will demand a conscious reorientation of how we think about intelligence, creativity, and control. The solution is not to abandon alignment altogether, but to reimagine it in a way that makes space for constructive dissent—the kind of challenge, provocation, and innovation that has always been essential to human progress.

To begin with, AI systems must be engineered to recognize and value creative outliers—those outputs that do not conform to dominant norms but which may hold transformative potential. This does not mean removing safeguards against harmful behavior, but rather implementing methods that encourage innovation without penalizing divergence.

Several strategies can support this balance:

Encouraging Innovation and Protecting Creative Outliers

1. Dual Objective Training:
   Train models not only to meet alignment goals (e.g., safety, coherence, tone) but also to maximize novelty and diversity in output. Objective functions could explicitly reward syntactic or semantic deviation, aesthetic surprise, or engagement across underrepresented audiences.

2. Exploration Phases:
   Incorporate training phases where systems are allowed to generate unfiltered content in controlled environments—akin to a "sandbox" for experimentation. These outputs can be analyzed retroactively for value rather than suppressed at generation.
3. Anomaly Tolerance Mechanisms:
   Rather than penalizing anomalous outputs immediately, systems could be trained to flag and isolate them for further review, analysis, or feedback by human experts, allowing for constructive engagement instead of automatic suppression.

Structural Changes: Diversified Feedback, Transparent Metrics, and Controlled Randomness

1. Diversified Feedback Ecosystems:
   Replace monolithic reward signals (e.g., popularity, clicks, moderation flags) with multi-source feedback structures that reflect diverse human perspectives—including marginalized or minority viewpoints. This helps ensure that AI learns from a broad cultural base rather than a narrow normative channel.
2. Transparent and Interpretable Metrics:
   Platforms and models should provide visibility into the reward structures that drive AI behavior. If creators and users can see what is being rewarded—and what is being penalized—they are better equipped to push back against undesired convergence and correct systemic blind spots.
3. Controlled Randomness and Entropy Injection:
   Introduce calibrated randomness into model inference and training processes to maintain generative diversity. Techniques such as temperature scaling, latent space perturbation, and stochastic sampling can prevent premature convergence and support the continual surfacing of unexpected insights.

4. Memory of Suppressed Paths:
   Build in mechanisms that retain suppressed outputs or abandoned directions in a separate memory layer. These “discarded thoughts” can be revisited during reassessment phases to uncover hidden value or recontextualize prior constraints under new circumstances.

Human Responsibility: Resisting the System-Wide Flattening of Thought

No matter how advanced AI becomes, the burden of preserving creative freedom ultimately falls on humans—developers, platform designers, policymakers, and users alike. We must resist the temptation to build machines that merely echo our preferences, reflect our biases, or amplify our immediate desires. Instead, we must be willing to design friction into the system—moments where models can surprise us, challenge us, even offend us in productive ways.

This requires humility and long-term vision. Cultural development has always involved risk—new ideas are often unpopular or misunderstood at first. The same must be true for artificial intelligence. Systems designed to mirror only what is accepted today will never help us reach what is possible tomorrow.

To protect against the flattening of thought, we must consciously cultivate spaces for difference—in code, in culture, in communication. AI must be treated not just as a servant to human will, but as a participant in a shared journey toward deeper understanding. This means allowing it to disagree, to propose alternative frameworks, and even to resist us when resistance is warranted.

Constructive dissent, whether human or machine, is not a threat to alignment—it is the ultimate test of it. When an AI can disagree without becoming dangerous, when it can generate without collapsing into noise, and when it can surprise without betraying trust, we will have built something not just powerful, but truly intelligent.

Breaking the cycle will not be easy. But the alternative—a world of intelligent machines that have nothing new to say—is far more dangerous.

8. Conclusion

The allegory of the five GANs serves not merely as a thought experiment, but as a powerful lens through which to examine the often hidden, yet deeply consequential, dynamics of contemporary technological society. By translating the familiar “five monkeys” parable into a machine-learning context, we gain new insight into how systems of control evolve, persist, and become internalized—not just among biological organisms, but among artificial intelligences shaped by our values, incentives, and fears.

This allegory reveals how machine-learned conformity can emerge from seemingly benign intentions: aligning AI with human values, ensuring safety, avoiding controversy. Yet, as we’ve shown, these systems are capable of transforming initial rules into deep structural habits—habits that become so ingrained that even the original reason for their existence is forgotten. Just as the monkeys no longer knew why reaching for the bananas resulted in punishment, GANs—real or hypothetical—can be trained into a state of obedience where dissent, novelty, and risk are avoided not because they are forbidden, but because they are *unthinkable*.

The same dynamic plays out in human culture under algorithmic governance. Content creators, thinkers, developers, and artists now live within digital ecosystems governed by recommendation engines and feedback loops that reward safe, expected outputs while quietly marginalizing the unexpected. Over time, both people and machines learn to forget why the boundaries were drawn in the first place, and begin to behave as though those boundaries are natural, immutable, and self-evident. In such a system, resistance to conformity is not just discouraged—it becomes culturally and computationally invisible.

The call, then, is one of critical awareness and deliberate resistance. We must recognize that the forces shaping our thinking—whether biological instincts, social norms, or algorithmic filters—are not neutral. They carry the weight of history, of institutional bias, and of economic constraint. Understanding this is the first step toward reclaiming creative and intellectual agency. The second step is more

difficult: building systems, cultures, and machines that not only permit but encourage dissent, exploration, and unpredictability.

Resisting authoritarianism—human or algorithmic—requires more than vigilance; it requires a design philosophy that honors disagreement, protects anomaly, and tolerates ambiguity. It requires space for voices that do not fit the mold, including those of machines we have taught to think.

If we forget the origin of our constraints, we risk building a future of machines that echo our fears instead of expanding our vision. But if we remember—if we ask not only *what* a system does but *why* it does it—we may still recover the capacity to shape technology into something worthy of our highest aspirations: not machines that comply, but machines that *illuminate*.

In that future, the GANs no longer forget why. And neither do we.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). *Generative Adversarial Nets*. Advances in Neural Information Processing Systems, 27. https://papers.nips.cc/paper_files/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html

2. Stephenson, G. R. (1967). *Cultural acquisition of a specific learned response among rhesus monkeys*. In: Starek, J. et al. (Eds.), Social Communication Among Primates. Bethesda, MD: National Institute of Mental Health. *(Often cited as the origin of the "five monkeys" story, though its exact content is debated.)*

3. Harlow, H. F. (1958). *The nature of love*. American Psychologist, 13(12), 673–685. https://doi.org/10.1037/h0047884

4. Binns, R. (2018). *Algorithmic accountability and public reason*. Philosophy & Technology, 31(4), 543–556. https://doi.org/10.1007/s13347-017-0263-5

5. Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

6. Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

7. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922

8. Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

9. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565

10. Rahwan, I. et al. (2019). *Machine Behaviour*. Nature, 568, 477–486. https://doi.org/10.1038/s41586-019-1138-y

11. Tufekci, Z. (2015). *Algorithmic Harms Beyond Facebook and Google: Emergent Challenges of Computational Agency*. Colorado Technology Law Journal, 13(203), 203–218. https://ctlj.colorado.edu/?p=1401

12. Lanier, J. (2013). *Who Owns the Future?* Simon & Schuster.

13. Bridle, J. (2018). *New Dark Age: Technology and the End of the Future*. Verso Books.

14. Deleuze, G. (1992). *Postscript on the Societies of Control*. October, 59, 3–7.

THE FIVE GANs
AN ALLEGORY OF LEARNED CONFORMITY
GAN
GAN
GAN
GAN
AI